RESEARCH PAPER

XCardioNet: An Explainable Cross-Modal Attention Hybrid Deep Learning Framework for Multimodal Cardiovascular Disorder Diagnosis

G. Amalorpavam[1] and N. Rajkumar[2]

[1]Research Scholar, Department of Computer and Information Science, Annamalai University, Annamalainagar, Tamilnadu, India. Email: amalphdcisau@gmail.com

[2]Assistant Professor [Deputed], Department of Computer Science, Manbumigu Dr Purachithalivar M.G.R Government Arts & Science College, Kattumannarkoil, Tamilnadu, India. Email: raju.prg@gmail.com

***Corresponding author: G. Amalorpavam, Research Scholar, Department of Computer and Information Science, Annamalai University, Annamalainagar, Tamilnadu, India.**

Email: amalphdcisau@gmail.com

ABSTRACT

The fast-paced rise in cardiovascular morbidity across the globe, therefore highlights the urgency for intelligent diagnostics systems application for early, accurate and clinically interpretable assessment of subclinical cardiac conditions. Although remarkable progress has been made using deep learning in computer-aided healthcare, none explored the integrated analysis of complementary information encoded by raw ECG signals and heterogeneous patient-specific clinical indicators while existing cardiovascular screening models almost solely focused on isolated unimodal information. Such a discretized learning framework typically struggles to achieve effective feature generalization, strong context-agnostic reasoning power, and clear decision making interpretability that is quite important in high-stakes pathology settings. In order to address these limitations, this work present XCardioNet: a next-generation explainable multimodal cardiovascular intelligence framework that jointly fuses ECG morphology learning using advanced representation methods with clinical risk-aware representation fusion for robust tri-class cardiac disorder identification. This architecture then consists of a dual-channel preprocessing engine that improves fidelity in the multimodal data: structured clinical records train a K-nearest neighbor imputer to fill in missing features alongside an Isolation Forest optimization step for outlier detection and feature normalization, while raw ECG recordings are optimized through a sequence of Butterworth spectral filtering, wavelet-domain denoising, adaptive R-peak identification and heartbeat-wise segmentation to ultimately preserve diagnostically salient waveform morphology. Then, a hierarchical classifier based on convolutional neural networks, Bi-directional long short-term memory and Transformer attention is used to learn localized beat characteristics, long-range temporal rhythm dependencies and global sequence semantics from segmented ECG patterns. A unified model of cardiovascular risk - Meanwhile, an attention-based clinical representation learner is able to extract underlying latent pathological signatures from high-dimensional clinical feature space. In contrast to traditional late fusion methods, the proposed framework embeds both modalities into a shared latent manifold and performs dynamic interdependency modelling through the multihead cross-modal attention fusion mechanism thereby supporting patient-to-signal and signal-to-patient information exchange for adaptive diagnosis. For the sake of clinical trustworthiness, further enhance the framework through SHapley Additive exPlanations-based feature attribution and attention heatmap visualization for interpretable reporting of modality contribution and decision relevance. Through extensive experimental validation, it is demonstrated that XCardioNet achieves a notable 99.14% diagnostic accuracy while greatly exceeding standalone ECG analyses, isolated clinical prediction and non-attentive multimodal baselines with substantially improved consistency in precision, recall and F1-score across all cardiovascular predicts. Additionally, a real-time intelligent graphical diagnostic prototype is created to confirm the translated feasibility of the proposed model in practical point-of-care settings. The presented findings establish that explainable cross-modal deep attention fusion can serve as a clinically scalable and computationally reliable paradigm for next-generation precision cardiovascular screening and early cardiac event prevention.

Keywords: Multimodal cardiovascular diagnosis; Electrocardiogram analysis; Clinical data fusion; Explainable artificial intelligence; Cross-modal attention; CNN–BiLSTM–Transformer; Deep learning in healthcare; SHAP interpretability

How to cite this article: Amalorpavam G, Rajkumar N. XCardioNet: An Explainable Cross-Modal Attention Hybrid Deep Learning Framework for Multimodal Cardiovascular Disorder Diagnosis. Int J Drug Deliv Technol. 2026;16(60s):61-78. DOI: 10.25258/ijddt.16.60s.7

Source of support: Nil.

Conflict of interest: None

1. Introduction

Cardiovascular diseases (CVDs) still represent one of the leading public health burden globally, being responsible for a large fraction of premature deaths and long-term disability in all age groups. To minimize the sudden fatal consequence and facilitate timely intervention, it is clinically essential in implementation to identify underlying functional or structural cardiac abnormalities (particularly

associated with progressing heart diseases) before developing acute episodes of significant myocardial injury due to an occlusion in coronary artery. Classic diagnosis methods are mainly based on manual electrocardiogram (ECG) interpretation, biochemical prediction and physician performed examination of patient clinical history. Yet, due to the amounts of data generated by patients increasing every day, inter-observer variability and relatively subtle electrophysiological manifestations manual diagnosis is slow, boring and an easy target for oversight. As a result, the incorporation of artificial intelligence (AI) into cardiovascular screening has recently emerged as an appropriate and obligatory thumbs up route of research for scalable and high precision clinical decision support [1].

In recent years, significant advancement has been made in the field of deep learning-assisted ECG interpretation: convolutional and recurrent architectures have shown a high level of ability to learn heartbeat morphology, rhythm irregularities and temporal dependencies from raw physiological waveforms. For instance, Zhang et al. [2] presented a fully explainable deep neural architecture for automatic ECG diagnosis, and they validated SHAP-guided model interpretability in cardiac abnormality detection. Likewise, Ahmad et al. [3] implemented a complex ECG segmentation and transfer deep learning model based on noise acquisition environments to complete the process of heart disease recognition, which is sensitive to waveform. Similarly, Poonkodi et al. introduced an attention-guided Bi-LSTM mechanism in [4], which effectively improves the retention of sequential ECG features for cardiovascular risk classification. Despite achieving impressive performance gains, these studies were inherently limited in their diagnostic scope by their reliance on ECG signals alone, thereby overlooking the vast amounts of pathological information contained within patient-level clinical records.

In reality, reliable diagnosis in practical cardiology is seldom derived from electrophysiological observations alone. Clinical determinants such as age, blood pressure, cholesterol profile, diabetic status, previous medical history and demographic risk properties complementarily elucidate by proxy disease severity or propensity and acute event susceptibility. Inspired by this clinical reality, recent studies have explored multimodal cardiovascular intelligence through abstracting from signals and metadata. Cao et al. [5] performed multimodal contrastive learning on ECGs and patient metadata and demonstrated that latent alignment of heterogeneous domains enhances the robustness of disease predictions. Mohsen et al. also confirmed that ECG-derived features possess more power for risk stratification on the patient level with a set of conventional risk factors than unimodal baselines [6]. Recently, Gupta et al. [7] developed a Conv-BiLSTM-Attention multimodal framework integrated with raw ECG and demographic factors for acute myocardial infarction detection and showed better generalizability in emergency environment. The combination of the findings confirms that multimodal fusion provides a more clinically coherent unifying diagnostic paradigm than isolated waveform interpretation.

However, the current state of the art is still limited when it comes to three major research challenges. First, numerous current multimodal investigations either employ shallow concatenation or late decision fusion; these approaches do not capture the deep interdependence between electrophysiologic dynamics and heterogeneous patient pathophysiology. Second, most cardiovascular predictive models depend on single-stage deep learners that are incapable of summarizing local heartbeat morphology, long-range temporal rhythm patterns, and global sequence semantics at once. Third, most of the current diagnostic architectures are still black-box if clinicians receive some feedback and interpretation they will not practice wide in high-risk medical settings. Recent explainable AI surveys state explicitly that clinically deployable cardiovascular intelligence should not only be more accurate but also need to include transparent modality contributions and trustworthy feature attribution [1], [8].

To address these limitations, this work proposes XCardioNet, an explainable cross-modal hybrid deep learning framework for intelligent multimodal cardiovascular disorder diagnosis from simultaneous ECG signals and structured clinical attributes. A strong dual preprocessing engine is first devised consist of KNN- driven missing value reconstruction, Isolation Forest-based outlier elimination and robust normalization for clinical records while ECG recordings are subjected to Butterworth bandpass filtering, wavelet-domain denoising, adaptive Rpeak localization and segmentation on heartbeat-wise basis in order to preserve morphology. Then a hierarchical ECG representation learner is designed to jointly summarize local electrophysiological textures, long-term rhythm continuity and contextual dependencies with convolutional neural networks (CNNs), bidirectional long short-term memory (BiLSTM) and Transformer attention. Simultaneously, the model uses multi-dimensional label-embedding sets generated by a transformer attention-guided clinical encoder to learn clinically meaningful representations of patient risk embeddings. In contrast to standard direct modality fusion approaches, the learned features are projected into a shared latent manifold that is adaptively fused by leveraging mutual interaction between physiological waveform behavior and patient centric pathological indicators via multihead cross-modal attention. Moreover, feature attribution via SHapley Additive

exPlanations (SHAP), cross-modal attention heatmap visualizations, and an intelligent graphical user interface (GUI) are included for transparency in decision support and clinical feasibility translation. Comprehensive experimental results demonstrate that the proposed framework outperforms (i) standalone ECG learning, (ii) isolated clinical prediction, and (iii) non-attentional multimodal baselines for a highly robust tri-class cardiovascular task with improved interpretability and deployability. Hence, the proposed study sets a clinically feasible route towards next-generation precision cardiac screening via explainable multimodal deep attention intelligence.

Contributions of the Proposed Work

The major novel contributions of this research are summarized as follows:

- A novel multimodal cardiovascular diagnostic framework, XCardioNet, is introduced to jointly exploit ECG electrophysiological behavior and heterogeneous patient clinical attributes for robust tri-class heart disorder prediction.
- A dual-stream preprocessing mechanism is developed that simultaneously performs clinical anomaly suppression and ECG morphology-preserving denoising/segmentation to enhance multimodal diagnostic fidelity.
- A hierarchical hybrid ECG encoder integrating CNN, BiLSTM, and Transformer attention is designed to capture localized heartbeat morphology, sequential rhythm dependency, and global contextual cardiac semantics.
- A patient risk-aware clinical attention encoder is constructed and dynamically fused with ECG embeddings through a multihead cross-modal attention mechanism, enabling deeper modality interaction than conventional concatenation-based fusion.
- An explainable AI module incorporating SHAP global attribution, class-wise feature relevance, and attention heatmap interpretation is embedded to improve clinician trust and diagnostic transparency.
- A real-time intelligent GUI-based clinical prototype is developed to demonstrate the translational applicability of the proposed architecture in practical point-of-care cardiovascular screening environments.

2. Related works

Advances in the use of artificial intelligence have recently changed automated cardiovascular screening by making machine-driven interpretation of electrophysiological signals, patient metadata, and heterogeneous multimodal clinical evidence possible. Existing literature in this field can be categorized into four main streams including ECG-based deep learning diagnosis, AI-based clinical risk prediction, multi-modal cardiovascular fusion intelligence and explainable reliable cardiac decision systems.

2.1 ECG-Centered Deep Learning for Cardiovascular Diagnosis

Electrocardiogram (ECG) analysis is still one of the most thoroughly studied directions in intelligent cardiovascular assessment owing to non-invasively and well electrophysiological relevance with the hidden cardiac disorder. Most of these recent data driven efforts aim to increase the ability of deep neural architectures to extract informative temporal and morphological features from raw ECG logs. Lee et al. [9] incorporated a cross-stage partial network with a cross-attention transformer to reproduce cardiac waveforms and found that attention-based temporal modelling renders ECG-based cardiovascular decision reliable. Chang et al. [10] proposed a practical deep learning solution for real-world ECG-based prediction of CAD and the need for coronary artery revascularization, demonstrating that large-scale electrophysiological AI screening can serve as efficient early warning assistance in clinical cardiology settings. Similarly, Dhandapani et al. [11] used a hybrid deep architecture on ECG signal images and robustly showed that learned deep visual electrophysiological descriptors outperform conventional handcrafted signal analysis. More recently, Lilhore et al. [12] developed an attention mechanism for the BiLSTM and this attention mechanism reinforced learning sequentially of the heartbeats dependency, by finding attention weights to each heartbeat within the input embedding time sequence, and using this modification in multiclass heart disease classification improved performance. Altogether, these studies demonstrate that ECG-driven deep representation learning is feasible; yet their diagnostic conclusions rely mainly on waveform manifestations alone and thus may not be robust against context incompleteness when diagnoses are made without requiring patient-level pathological risk indicators to be jointly examined.

2.2 AI-Assisted Clinical and Structured Cardiovascular Risk Prediction

In parallel with the ECG-based analysis, a number of studies have explored machine learning and deep neural approaches to assess the contribution of structured clinical attributes in identifying cardiovascular disorders. By leveraging heterogeneous patient risk records, Fitriyani et al. [13] developed a powerful heart disease prediction model for clinical decision support and showed that intelligent tabular learning can make potential improvement to consistency of disease screening. Building on this idea, a complete systematic review on AI-based cardiovascular risk prediction by Cai et al. [14] came to the conclusion that demographic profiles, laboratory parameters and medical history provide important early-stage indicators of patient

risk. According to Bartusik-Aebisher et al. [15], AI-assisted ECG interpretation combined with preventive clinical intelligence has the potential to advance non-invasive cardiac diagnostics for the next generation. While these clinical prediction systems can offer useful patient-centered pathological information, they generally fail to provide direct electrophysiological information about dynamic cardiac rhythm disturbances that could be of key diagnostic importance, thus limiting sensitivity in cases where abnormalities of the waveforms are diagnostically important.

2.3 Multimodal Fusion Intelligence in Cardiovascular Screening

Realizing the constraints of unimodal learning, recent studies have progressively shifted towards multimodal cardiovascular diagnosis by combining ECG with supplementary clinical or imaging modalities. A multimodal fusion framework of ECG and chest X-ray radiomics was proposed by Sun et al. [16], and the nomogram analysis based on machine learning showed that heterogeneous modality interaction can significantly enhance diagnostic robustness against single-source evidence. Soto et al. [17] using latent inter-domain representation coupling for multimodal deep learning and its application to multi-modal data fusion improved the left ventricular hypertrophy identification by jointly modelling electrical and structural cardiac information. A unified patient-centric decision modelling for multimodal artificial intelligence through the lens of precision medicine was provided in [18] to give a wider perspective on how it can advance cardiovascular disease management by Yang et al. Accordingly, Archana and Shashikala [19] proposed a domain-based hybrid multimodal heart disease prediction framework that utilize multimodal medical image inputs and feature fusion strategies, through the demonstration of heterogeneous information aggregation outperforming homogeneous information for complex cardiac diagnosis. Ramos-Zaga [20] reviewed the advanced framework of future multimodal diagnostic approaches in cardiovascular disease and suggested that the AI-enabled screening systems should now shift from standalone bio signal processing to synchronized multimodal clinical reasoning. While these studies provide clear evidence that multi-modal learning enhances the comprehensiveness of analysis, many still follow traditional approaches such as feature concatenation or shallow fusion, which are improperly synthesized and do not model deep semantic interactions across different modalities of medical evidence.

2.4 Cross-Attention and Explainable Cardiovascular Intelligence

A more relevant and very prominent trend in the literature is about all attention-guided, explainable cardiovascular AI systems which can enhance diagnostic performance without compromising clinician trust. Finally, Majhi and Kashyap [21] presented an AI-based machine learning approach for ECG-based heart disease prediction with explainability, providing preliminary peer-to-peer evidence that interpretable feature contribution analysis could significantly increase the transparency of a classification model and clinician's confidence. Qu et al. MAF-Net, a multimodal cross-attention based fusion network for heart disease classification was introduced in [22], and the authors demonstrated that dynamic cross-modal interaction learning is more advantageous than static fusion baselines. Dong and Xie [23] adapted their multimodal cross-attention CNN-LSTM known as MC-CNN-LSTM for arrhythmia classification and clinical diagnostic assistance, demonstrating that explicit modality dependency modelling can greatly boost heterogeneous cardiac decision making. In summary, these studies suggest cross-attention learning and explainable inference is likely to be a fruitful approach toward clinically trustworthy cardiovascular AI development. However, the current models handle either multimodal fusion or explainability as decoupled improvements instead of jointly optimizing these components, and very few systems consolidate hierarchical ECG feature learning, patient clinical risk encoding, dynamic cross-modal attention reasoning and transparent diagnostic attribution into the same end-to-end system.

2.5 Research Gap Identified

As highlighted in the reviewed studies [9] - [23], it is evident that there is no single joint cardiovascular diagnostic system that integrates hierarchical ECG representation learning, patient risk-aware clinical embedding, multihead cross-modal attention fusion and dual-level explainable interpretation over a deployable real-time screening environment. ECG-only models are limited by pathological incompleteness, clinical-only models lack physiological responsiveness, and current multimodal systems use weak semantic fusion but do not fully model the dynamic interaction of electrophysiological and patient-specific diagnostic information. Additionally, even though explainable AI is a trending topic nowadays, multimodal interpretability with integrated transparency through feature attribution and cross-attention reasoning has not been adequately studied. These account for major methodological needs of the proposed XCardioNet framework yet to be resolved.

3. Proposed Methodology

This Section presents the proposed framework for intelligent tri-class cardiovascular disorder diagnosis as an Explainable Cross-Modal Hybrid Deep Learning Approach. The framework is structured as an end-to-end multimodal processing pipeline, which methodically combines data preprocessing, separate deep feature learning with

modal safety margin, adaptive fusion of multimodal representations for interpretable fusion-based multiview supervised clustering analysis, and real-time diagnostic deployment. First, the dedicated data preprocessing methods in this study performed noise reduction, missing value treatments, outlier detection and correction (ODC), as well as normalization of heterogeneous multi-criteria diagnostic inputs on both raw ECG signals and structured clinical records. Then, it extracts discriminative latent features in a modality-independent manner by using a hierarchical CNN–BiLSTM–Transformer encoder to learn local and temporal ECG characteristics from RRIs, while at the same time novel attention guided clinical encoder learns patient specific pathological risk information provided as external covariates. The fused clinical evidence is then projected in a shared latent space and dynamically combined with the multimodal embeddings using multihead cross-modal attention to capture the dependency across these two domains. SHAP-based feature attribution and attention heatmap analysis provide interpretability to ensure transparent clinical decision support. Finally, as a practical applicability of the system for real-time cardiovascular screening, an intelligent intuitive graphical user interface is developed. The complete block diagram of the proposed framework is shown in Figure 1, and the Table 1 gives details about the layer-wise architectural settings.

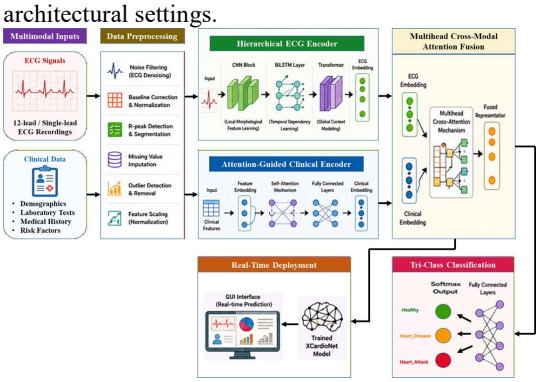


Figure 1: The Block Diagram of the Proposed XCardioNet Framework

Table 1. Layer-wise Architecture of the Proposed XCardioNet Framework

Layer Name	Type	Output Shape	Description
Input Data	Multimodal Diagnostic Input	ECG: (N,360), Clinical: (M,)	Heartbeat ECG segments and structured patient clinical records
ECG Preprocessing	Filtering + Denoising + Segmentation	(N,360)	Butterworth filtering, wavelet denoising, R-peak detection and heartbeat extraction
Clinical Preprocessing	Imputation + Outlier Removal + Scaling	(M,)	KNN imputation, Isolation Forest cleaning and Robust normalization
ECG Encoder - CNN	Conv1D + MaxPooling	(N,64)	Extracts local morphological heartbeat patterns
ECG Encoder - BiLSTM	Bidirectional LSTM	(N,128)	Learns sequential temporal rhythm dependencies
ECG Encoder - Transformer	Multihead Self-Attention Encoder	(128)	Captures long-range global electrophysiological context
Clinical Encoder	Dense + Self-Attention Layers	(128)	Generates patient-centric pathological risk embedding
Latent Projection	Fully Connected Mapping	ECG: (128), Clinical: (128)	Projects both modalities into common latent manifold
Cross-Modal Fusion	Multihead Cross-Attention	(128)	Learns adaptive dependency between ECG and clinical embeddings
Classifier	Dense + ReLU + Dropout + Softmax	(3)	Performs tri-class cardiovascular disorder prediction
Explainability Module	SHAP + Attention Heatmap	Feature-wise	Provides global, class-wise and fusion-level interpretation
Deployment Module	GUI-Based Real-Time Inference	Prediction Panel	Supports live diagnostic screening and clinician decision assistance

3.1 Multimodal Data Preprocessing

Multimodal inputs obtained from heterogeneous medical sources are often in bad quality and unstructured nature, hence the accuracy and trustworthiness of cardiovascular diagnosis relies heavily on how these multimodal inputs can be extracted consistently. In the XCardioNet framework proposed here, that jointly leverages structured clinical records and raw electrocardiogram (ECG) waveforms as input sources, modality-specific noise suppression, irregularity removal, and representation quality enhancement strategies need to be applied. Therefore, a two-stream multimodal preprocessing pipeline including clinical data refinement and ECG signal conditioning is proposed as shown in Figure 1.

3.1.1 Clinical Data Refinement

The clinical records are raw and consist of heterogeneous attributes including demographic, vital signs, laboratory observations and historical cardiac risk factors. Let the original clinical dataset be represented as

$$C = \{c_i\}_{i=1}^{N}, \quad c_i \in \mathbb{R}^M \tag{1}$$

Where N denotes the total number of patient samples and M represents the number of clinical attributes.

(a) Missing Value Imputation

Many attributes exhibit missing observations because of incomplete patient entries. K-nearest neighbor (KNN) imputation is utilized to "fill in" these missing values. For a missing attribute c_{ij}, its imputed estimate is computed as

$$\hat{c}_{ij} = \frac{1}{K}\sum_{k \in N_K(i)} c_{kj} \tag{2}$$

Where $N_K(i)$ denotes the set of K nearest neighboring patients of sample i based on Euclidean similarity in the observed feature space.

(b) Outlier Detection and Removal

Abnormal physiological measurements in clinical datasets are problematic due to acquisition errors or consistent entries that are rare and inconsistent. For this, outlier filtering based on Isolation Forest is used to mitigate these anomalies. For each patient sample c_i, an anomaly score $A(c_i)$ is determined as

$$A(c_i) = 2^{-\frac{E\left(h(c_i)\right)}{d(N)}} \tag{3}$$

Where $E\left(h(c_i)\right)$ is the expected path length of the sample in the random isolation trees and $d(N)$ is the average path normalization constant. Samples with $A(c_i) > T$ are treated as abnormal and removed from the clinical training pool.

(c) Robust Feature Normalization

In particular, clinical variables are measured in different units; therefore, they all go through RobustScaler normalization to reduce the influence of a few extreme values of residuals. The normalized clinical feature c_{ij}^* is computed as

$$c_{ij}^* = \frac{c_{ij} - Q_2(c_j)}{Q_3(c_j) - Q_1(c_j)} \tag{4}$$

Where Q_1, Q_2, and Q_3 represent the first quartile, median, and third quartile of the j^{th} feature, respectively. Thus, the refined clinical feature matrix is obtained as

$$C^* \in \mathbb{R}^{N \times M} \tag{5}$$

which provides a stable patient-centric pathological representation for downstream learning.

3.1.2 ECG Signal Conditioning

The ECG raw recordings are very prone to baseline wander, motion artefacts, muscle noise and acquisition disturbances. This motivates the development of a 4-step ECG preprocessing pipeline that preserves clinically meaningful electrophysiological morphology. Let the raw ECG waveform of the i^{th} patient be denoted as

$$x_i = \{x_i(t)\}_{t=1}^{T} \tag{6}$$

Where T is the signal length.

(a) Bandpass Noise Filtering

A fourth-order Butterworth bandpass filter performs in the frequency range $0.5\ \text{Hz} - 40\ \text{Hz}$ to eliminate low-frequency baseline drift and high-frequency sensor noise. The filtered signal is expressed as

$$x_i^{(1)}(t) = B\big(x_i(t)\big) \tag{7}$$

Where $B(\cdot)$ denotes the Butterworth filtering operator.

(b) Wavelet Denoising

Discrete wavelet threshold denoising is applied to reduce more of the non-stationary transient artifacts while preserving QRS morphology:

$$x_i^{(2)}(t) = W^{-1}\left(\Theta\left(W\left(x_i^{(1)}(t)\right)\right)\right) \tag{8}$$

Where W and W^{-1} denote forward and inverse wavelet transforms, and $\Theta(\cdot)$ is the soft-thresholding operator.

(c) Signal Normalization

The denoised ECG signal is standardized to zero mean and unit variance:

$$x_i^{(3)}(t) = \frac{x_i^{(2)}(t) - \mu_i}{\sigma_i + \epsilon} \tag{9}$$

Where μ_i and σ_i are the mean and standard deviation of the signal, and ϵ is a small constant to avoid numerical instability.

(d) R-Peak Detection and Heartbeat Segmentation

Since heartbeat-centered morphology is more informative than continuous raw traces, adaptive R-

peak localization is performed: $P_i = \{p_1, p_2, \dots, p_n\}$ where P_i denotes the detected R-peak positions.
For each detected peak p_k, a fixed window of length $2w$ is extracted to generate heartbeat segments:

$$s_k = x_i^{(3)}[p_k - w : p_k + w] \quad (10)$$

Thus, the segmented ECG sample set becomes

$$S_i = \{s_k\}_{k=1}^{n}, \quad s_k \in \mathbb{R}^{2w} \quad (11)$$

which preserves localized P-QRS-T morphological patterns for deep electrophysiological encoding.

3.1.3 Preprocessed Multimodal Output

After the above refinement procedures, the final preprocessed multimodal input pair can be represented as

$$I^* = \{S_i, C_i^*\} \quad (12)$$

Where S_i is set of normalized heartbeat ECG segments, and C_i^* is cleaned and scaled clinical feature vector.
This dual-stream modality-consistent representation ensures that the subsequent XCardioNet encoders learn from diagnostically reliable physiological and pathological evidence with reduced noise interference and enhanced cross-modal compatibility.

3.2 Hierarchical ECG Representation Learning

Electrocardiogram signals demonstrate rich electrophysiological patterns over which diagnostically relevant information is spread across local morphology of heartbeats, sequential continuity of rhythm and long-range temporal context. Due to heterogeneous characteristics of these signals, conventional shallow feature extractors often cannot simultaneously capture the key information from electrocardiograms and triaxial accelerations for robust downstream cardiovascular prediction. To overcome this limitation, this work presents a new XCardioNet framework that utilizes a hierarchical ECG representation learning module to effectively model the local, mid-level and global electrophysiological semantics of ECGs via the collaborative embedding of CNN-BiLSTM-Transformer attention encoding. Figure 2 shows the entire stream of ECG encoding.

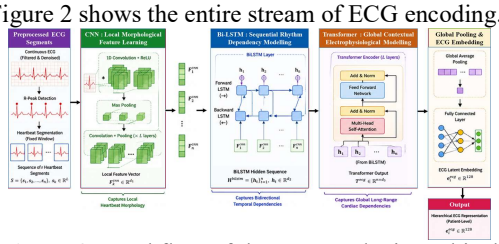


Figure 2. Workflow of the Proposed Hierarchical ECG Representation Learning Module

Let the preprocessed heartbeat segment set of the i^{th} patient obtained from equation (11) be denoted as $S_i = \{s_k\}_{k=1}^{n}, \quad s_k \in \mathbb{R}^L$, where n is the number of segmented beats and L represents the fixed segment length.

3.2.1 Local Morphological Feature Learning using CNN

The segments representing each heartbeat contain clinically relevant local structures like P-wave amplitude variation, deformation of QRS complex, ST-segment shift and T-wave irregularity. The first step is to retain these fine-grained morphological properties using one-dimensional convolutional learning.
For an input segment s_k, the output of the m^{th} convolutional filter is defined as

$$f_k^{(m)} = \sigma(W_m * s_k + b_m) \quad (13)$$

Where W_m denotes the learnable convolution kernel, $*$ indicates one-dimensional convolution, b_m is the bias term, $\sigma(\cdot)$ represents the nonlinear ReLU activation.
Subsequently, max-pooling is applied to suppress redundant activations and preserve dominant morphological responses:

$$p_k^{(m)} = max\left(f_k^{(m)}\right) \quad (14)$$

After passing through successive convolution–pooling blocks, the local ECG representation can be expressed as

$$F_k^{cnn} \in \mathbb{R}^{d_1} \quad (15)$$

Where d_1 denotes the learned local morphological embedding dimension. This CNN stage enables robust extraction of heartbeat shape descriptors while reducing minor signal perturbations.

3.2.2 Sequential Rhythm Dependency Modeling using BiLSTM

While CNN adeptly learns disease-specific morphology from spatiotemporal local snapshots, it overlooks cardiac dysfunction occurring as a result of temporal rhythm transitions between consecutive heartbeats. The CNN local features are sequentially processed by a BiLSTM layer in order to model the bidirectional dependency.
Let the ordered CNN feature sequence be

$$F^{cnn} = \{F_1^{cnn}, F_2^{cnn}, \dots, F_n^{cnn}\} \quad (16)$$

The forward and backward LSTM hidden states are computed as

$$\overrightarrow{h_t} = LSTM_{fwd}\left(F_t^{cnn}, \overrightarrow{h_{t-1}}\right), \qquad \overleftarrow{h_t} = LSTM_{bwd}\left(F_t^{cnn}, \overleftarrow{h_{t+1}}\right) \quad (17)$$

The bidirectional temporal representation is then obtained by concatenation:

$$h_t = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right] \quad (18)$$

Thus, the complete rhythm-aware ECG sequence becomes

$$H^{bilstm} = \{h_t\}_{t=1}^{n}, \quad h_t \in \mathbb{R}^{d_2}$$

(19)

Where d_2 is the BiLSTM hidden dimension. This produces a form of embedding that accounts for both preceding and succeeding beat dependencies, enabling the model to detect rhythm irregularities, intermittent anomalies, and continuity based pathological transitions.

3.2.3 Global Contextual Electrophysiological Modeling using Transformer

BiLSTM is a promising model to capture local temporal continuity, but it may not sufficiently represent long-range cardiac relations distributed over distant heartbeat positions. Thus, the work utilizes a Transformer encoder for global self-attentive contextual interaction in an ECG sequence. Given the BiLSTM feature matrix H^{bilstm}, query, key, and value projections are computed as

$$Q = H^{bilstm} W_Q, \quad K = H^{bilstm} W_K, \quad V = H^{bilstm} W_V \tag{20}$$

Where W_Q, W_K, and W_V are learnable projection matrices. The scaled dot-product self-attention is then formulated as

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{21}$$

For multihead attention with H heads:

$$MHA(Q,K,V) = Concat(head_1, \dots, head_H)W_O \tag{22}$$

where each head independently learns distinct global cardiac dependencies. The Transformer-enhanced contextual representation is thus obtained as

$$T^{ecg} \in \mathbb{R}^{n \times d_3}$$

(23)

which contains globally aware electrophysiological interactions across the heartbeat sequence.

3.2.4 Global Pooling and ECG Latent Embedding

To generate a compact patient-level ECG representation suitable for multimodal fusion, average global pooling is performed over all Transformer outputs:

$$z_i^{ecg} = \frac{1}{n}\sum_{t=1}^{n} T_t^{ecg}$$

(24)

The pooled vector is then projected through a fully connected embedding layer:

$$e_i^{ecg} = W_e z_i^{ecg} + b_e$$

(25)

Where $e_i^{ecg} \in \mathbb{R}^{128}$ represents the final latent ECG embedding learned for the i^{th} patient.

This hierarchical encoding strategy ensures that the proposed framework simultaneously captures local heartbeat morphology via CNN, bidirectional temporal rhythm continuity via BiLSTM, and global long-range cardiac dependency via Transformer. As a result, the generated ECG embedding becomes significantly more discriminative and clinically informative than single-stage electrophysiological feature extractors.

3.3 Attention-Guided Clinical Feature Encoding

Apart from the abnormalities of electrophysiological waveforms, cardiology diagnosis relies heavily on patient-centered pathological markers—age, blood pressure, laboratory measurements as well as comorbidity history and various clinical risk factors. The heterogeneous structured attributes together provide corroborating diagnostic evidence that cannot be inferred directly from ECG morphology alone. However, not all clinical variables play an equally important role in disease manifestation and naïve equal-weight learning can wash out diagnostically informative signals. To this end, an attention-guided clinical feature encoding module is introduced in the proposed XCardioNet framework to adaptively highlight high-risk pathological features while reducing clinically less informative components.

Let the preprocessed clinical feature vector of the i^{th} patient obtained from equation (5) be denoted as $c_i^* = [c_{i1}, c_{i2}, \dots, c_{iM}]^T \in \mathbb{R}^M$, where M represents the number of normalized clinical variables.

3.3.1 Dense Clinical Feature Projection

Given that the raw clinical attributes are in heterogeneous measurement scales and semantic spaces, these attributes go through an initial nonlinear dense projection to obtain a corresponding latent pathological representation. The first hidden projection is formulated as

$$h_i^{(1)} = \phi(W_1 c_i^* + b_1)$$

(26)

Where W_1 is the learnable weight matrix, b_1 is the bias vector, and $\phi(\cdot)$ denotes the ReLU activation.

A second dense transformation is then applied:

$$h_i^{(2)} = \phi\left(W_2 h_i^{(1)} + b_2\right) \tag{27}$$

such that $h_i^{(2)} \in \mathbb{R}^{128}$. This operation allows the model to learn nonlinear interactions between demographic, biochemical and physiological risk features.

3.3.2 Attention-Based Risk Importance Weighting

Despite their generative projecting as latent clinical interpretable in our particular data, not all hidden components contribute equally to the candidate final cardiovascular diagnosis. To this end, an adaptive attention weighting mechanism is introduced to measure the relative significance of each latent pathological feature. The attention score vector is computed as

$$u_i = tanh\left(W_a h_i^{(2)} + b_a\right) \tag{28}$$

Where W_a and b_a are trainable attention parameters. The normalized clinical attention weights are then obtained using the Softmax function:

$$\alpha_i = Softmax(W_s u_i + b_s)$$

(29)

Where $\alpha_i = [\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{i128}]$ and $\sum_{j=1}^{128} \alpha_{ij} = 1$. These weights directly reveal the severity of each latent clinical feature in terms of pathological contribution.

3.3.3 Risk-Aware Clinical Embedding Generation

The final attention-guided clinical representation is obtained by element-wise feature reweighting:

$$z_i^{clin} = \alpha_i \odot h_i^{(2)}$$

(30)

Where $\odot$ denotes Hadamard (element-wise) multiplication.

Thus, $z_i^{clin} \in \mathbb{R}^{128}$ represents the patient-specific clinical embedding that selectively emphasizes dominant cardiovascular risk indicators while attenuating weak or redundant variables. To further stabilize the latent representation for multimodal interaction, the weighted vector is passed to the downstream embedding interface as $e_i^{clin} = z_i^{clin}$, which serves as the clinical branch output for cross-modal fusion.

The proposed attention-guided clinical encoder offers Nonlinear pathological interaction learning, Adaptive risk prioritization, and Fusion-ready latent representation. Therefore, rather than uniformly handling all patient variables as in existing works, the proposed module learns a strong foundation of risk-aware pathological representation for differentiated and improved patient-centric cardiovascular reasoning through inside the XCardioNet framework.

3.4 Multihead Cross-Modal Attention Fusion

While independent ECG and clinical encoders can learn modally-specific cardiovascular signatures, direct use of these isolated embeddings is inadequate for clinically sensible diagnosis given that cardiac abnormalities are seldom expressed through one source of in evidence Electrophysiological waveform disturbances and patient-specific pathological risk factors interact largely in a complementary, often interdependent manner in practical cardiovascular screening. Thus, such a latent diagnostic dependence cannot be satisfactorily captured by either naive feature concatenation or late averaging. To overcome this limitation, this work proposes a novel XCardioNet framework, which integrates a carefully designed Multihead Cross-Modal Attention Fusion module to dynamically learn the contextual interplay between ECG-derived electrophysiological semantics and clinically encoded pathological risk representations prior to tri-class final decision making.

Let the patient-level latent embeddings be represented as $e_i^{ecg} \in \mathbb{R}^{128}$, $e_i^{clin} \in \mathbb{R}^{128}$ for the i^{th} patient sample.

3.4.1 Common Latent Projection

While both modalities exist in equal-dimensional space, their internal semantic distributions are still modality-specific. And therefore, a first fully connected projection is needed to pull them together into their common latent interaction manifold.

$$q_i = W_q e_i^{ecg} + b_q$$

(31)

$$k_i = W_k e_i^{clin} + b_k$$

(32)

$$v_i = W_v e_i^{clin} + b_v$$

(33)

Where q_i denotes ECG-guided query representation, k_i denotes clinical key representation, v_i denotes clinical value representation, and W_q, W_k, W_v are learnable projection matrices. This projection allows the ECG modality to selectively attend toward clinically relevant pathological dimensions.

3.4.2 Cross-Modal Attention Weight Computation

The dependency between electrophysiological and clinical evidence is modeled with scaled dot-product cross-attention:

$$\beta_i = Softmax\left(\frac{q_i k_i^T}{\sqrt{d}}\right)$$

(34)

Where d denotes the latent embedding dimension. The attention coefficient β indicates the importance of each clinical latent component to the ECG electrophysiological query. The ECG-conditioned clinical response vector is then formulated as $r_i = \beta_i v_i$. Therefore, rather than concatenating both modalities directly, the fusion mechanism selectively filters the clinical representation by ECG driven pathological relevance.

3.4.3 Multihead Interaction Learning

Single-head attention can only be focused on one modality correlation, but in cardiovascular diagnosis, hidden representation interacting with other modalities is the key problem to solve because rhythm–age dependence interaction and waveform–blood pressure correlation and ischemic morphology–biochemical marker relation exists.

For the h^{th} head:

$$head_h = Attention\left(q_i^{(h)}, k_i^{(h)}, v_i^{(h)}\right)$$

(35)

Where $Attention(q, k, v) = Softmax\left(\frac{qk^T}{\sqrt{d_h}}\right)v$.

The outputs from all heads are concatenated $R_i = Concat(head_1, head_2, \dots, head_H)$ and projected through an output transformation $f_i = W_o R_i + b_o$. Where $f_i \in \mathbb{R}^{128}$ represents the unified fused multimodal cardiovascular embedding. This multihead strategy allows for synchronous learning of different cross-modal pathological interactions between ECG and clinical domains.

3.4.4 Adaptive Residual Fusion Stabilization

Residual fusion reinforcement is employed to employ dynamic learning of interaction while preserving original modality semantics:

$$\tilde{f}_i = f_i + e_i^{ecg} \quad (36)$$

This residual enrichment guarantees that the last fused embedding encodes inherent electrophysiological morphology knowledge, clinically attended pathological evidence, and cross-modal diagnostic dependency. Thus, $\tilde{f}_i \in \mathbb{R}^{128}$ becomes the final multimodal latent representation forwarded to the classification module.

The proposed Multihead Cross-Modal Attention Fusion module allows an embedding that is much more discriminative and clinically meaningful than a static fusion, thus acting as the central intelligence of XCardioNet framework.

3.5 Tri-Class Cardiovascular Classification

The Multihead Cross-Modal Attention Fusion module generates a fused latent representation that incorporates synchronized electrophysiological and pathological cardiovascular evidence after adaptive multimodal interaction learning. But this high dimension feature embedding needs to be converted into a clinically interpretable diagnostic decision (normal, moderate pathological disorder, acute high-risk cardiac event). To this end, the proposed XCardioNet framework applies a fully connected nonlinear classification head combined with Softmax probabilistic inference to contextually classify tri-class cardiovascular disorders. Let the final fused multimodal embedding be denoted as $\tilde{f}_i \in \mathbb{R}^{128}$ for the i^{th} patient sample.

3.5.1 Dense Nonlinear Decision Transformation

Then, applied a dense hidden layer to convert the latent fused representation into class-separable diagnostic features.

$$h_i^{cls} = \phi\left(W_c \tilde{f}_i + b_c\right) \quad (37)$$

Where W_c denotes the classifier weight matrix, b_c is the bias vector, and $\phi(\cdot)$ represents the ReLU nonlinear activation.

Thus, $h_i^{cls} \in \mathbb{R}^{64}$. It compresses the multimodal embedding, while retaining salient cardiovascular decision semantics.

Neural networks can overfit due to redundancy in the latents - so dropout regularization is introduced:

$$\hat{h}_i^{cls} = Dropout(h_i^{cls}) \quad (38)$$

where a fraction of hidden activations are randomly deactivated during training.

3.5.2 Softmax Probability Estimation

The regularized hidden representation is then mapped into a three-dimensional output decision vector:

$$o_i = W_o \hat{h}_i^{cls} + b_o \quad (39)$$

Where $o_i = [o_{i1}, o_{i2}, o_{i3}]$ corresponding to the three cardiovascular classes.

To obtain normalized posterior probabilities, the Softmax activation is employed:

$$P(y = j|x_i) = \frac{e^{o_{ij}}}{\sum_{k=1}^{3} e^{o_{ik}}} \quad \text{for } j \in \{1,2,3\}. \quad (40)$$

Accordingly, the predicted class probability vector becomes

$$P_i = [p_i^H, p_i^D, p_i^A] \quad (41)$$

Where p_i^H is probability of Healthy, p_i^D is probability of Heart_Disease, and p_i^A is probability of Heart_Attack.

3.5.3 Final Decision Rule

The final cardiovascular diagnosis is assigned based on maximum posterior probability:

$$\hat{y}_i = \arg \max_{j \in \{H,D,A\}} P_{ij} \quad (42)$$

Where $\hat{y}_i \in \{\text{Healthy}, \text{Heart\_Disease}, \text{Heart\_Attack}\}$.

The framework thus proposed clinically interpretable tri-level stratification: Healthy is absence of significant cardiovascular abnormality, Heart_Disease is presence of moderate pathological cardiac dysfunction that requires clinical attention, and Heart_Attack is acute high-risk myocardial abnormality requiring urgent intervention.

3.5.4 Optimization Objective

To ensure robust multiclass decision learning, the classification head is optimized using categorical cross-entropy loss:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{N} \sum_{j=1}^{3} y_{ij} log\left(P(y = j|x_i)\right) \quad (43)$$

Where y_{ij} is the ground truth label indicator, and $P(y = j|x_i)$ is the predicted posterior probability.

Minimization of $\mathcal{L}_{cls}$ enables the classifier to maximize diagnostic separation among healthy subjects, chronic heart disorder patients, and acute heart attack cases.

The tri-class decision module provides Clinically meaningful risk stratification, as opposed to a simple normal/abnormal prediction; Probability-driven confidence estimation for physician support; and Robust nonlinear separation of multimodal latent cardiovascular patterns. As a result, the classification module converts the fused cross-modal intelligence of XCardioNet into an interpretable patient-level cardiovascular diagnostic output for downstream explainability and deployable in real-time systems.

3.6 Explainable Decision Interpretation

While deep multimodal learning yields strong cardiovascular prediction capability, the clinical usability of such systems is constrained if the diagnostic decisions are produced by opaque black-box inference. Physicians in high-risk medical environments need accurate classification outcomes,

as well transparent evidence of the physiological and pathological factors responsible for any particular decision. To increase clinical trustworthiness and diagnostic accountability, the proposed XCardioNet establishes a two-level Explainable Decision Interpretation (EDI) module consisting of SHAP-based activation explanation with clinical feature attribution and cross-modal attention importance visualization. Let the final classification decision for the i^{th} patient be denoted as $\hat{y}_i = f(S_i, C_i^*)$, where $f(\cdot)$ represents the complete XCardioNet inference function.

3.6.1 SHAP-Based Global and Class-Wise Clinical Feature Attribution

SHapley Additive exPlanations (SHAP) are then applied to assess the contribution of each clinical variable to the final cardiovascular decision. SHAP calculates the marginal contribution of a feature by taking an average its effect over every possible contribution of features subsets. For the j^{th} clinical feature, the SHAP value is defined as

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{J\}) - f(S)] \quad (44)$$

Where F denotes the full clinical feature set, S represents a subset excluding feature j, $f(S)$ is the model prediction using subset S.

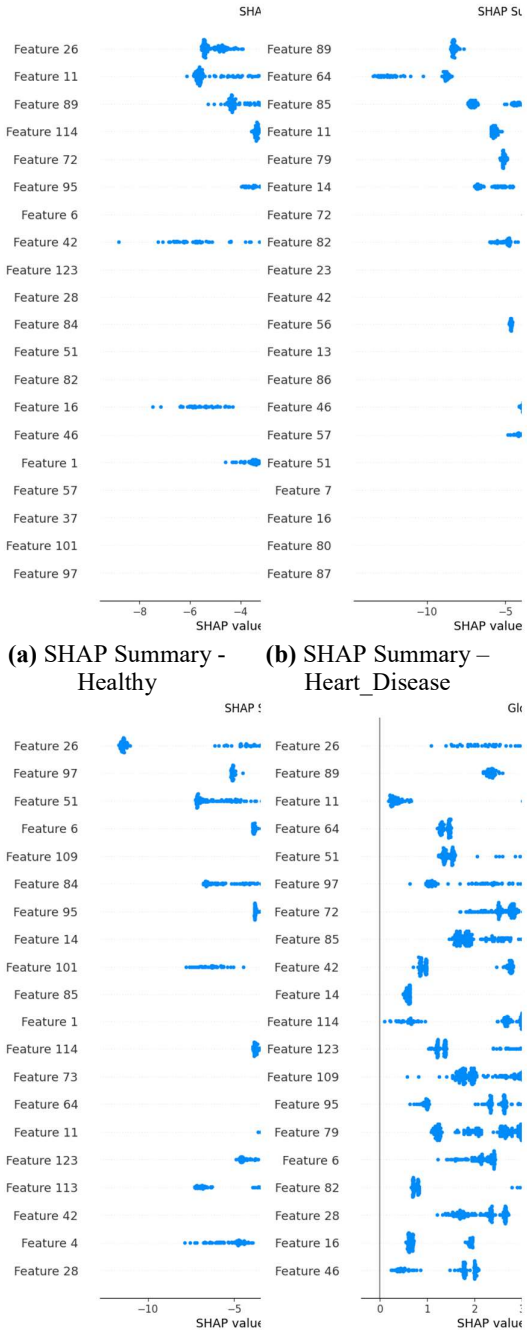
SHAP score ϕ_j obtained will either push the predicted cardiovascular class higher or lower, corresponding to its respective clinical factor. For global feature relevance across all N samples, the mean absolute SHAP importance is computed as

$$I_j^{global} = \frac{1}{N} \sum_{i=1}^{N} |\phi_{ij}| \quad (45)$$

Where ϕ_{ij} denotes the SHAP contribution of feature j for patient i. Similarly, class-specific diagnostic attribution is evaluated as

$$I_j^{(c)} = \frac{1}{N_c} \sum_{i \in c} |\phi_{ij}^{(c)}| \quad (46)$$

Where N_c denotes the number of samples belonging to class c. This enables the framework to explicitly recognize risk factors that are dominant globally regarding cardiovascular health and class-dependent pathological features driving Healthy, Heart_Disease, and Heart_Attack predictions. Figure 3 visually presents the class-wise and global SHAP contributions of each feature.



(a) SHAP Summary - Healthy **(b)** SHAP Summary – Heart_Disease

(c) SHAP Summary – Heart_Attack **(d)** Global SHAP Summary

Figure 3. SHAP-based visualization of class-wise and global feature contribution patterns

3.6.2 Cross-Modal Attention Relevance Analysis

This work goes further than SHAP, which explains the significance of clinical variabless but not how this interacts with ECG electrophysiological information through multimodal fusion based on the underlying clinical pathology. Therefore, the attention weights generated in Section 3.4 are further utilized to visualize cross-modal decision

dependency. Let the multihead attention coefficients for the i^{th} sample be represented as $\beta_i^{(h)}$, $h = 1,2,\dots,H$. The average fusion relevance map is computed as

$$\bar{\beta}_i = \frac{1}{H}\sum_{h=1}^{H}\beta_i^{(h)} \quad (47)$$

Where $\bar{\beta}_i$ denotes the mean ECG-to-clinical interaction intensity. Across the entire validation cohort, the global cross-modal attention matrix is obtained as

$$A = \frac{1}{N}\sum_{i=1}^{N}\bar{\beta}_i \quad (48)$$

This transformation is visualized through an attention heatmap which shows latent clinical dimensions most strongly activated by final decision making ECG electrophysiological queries. Therefore, the attention relevance map can clearly reveal which clinical features from ECG data are more important, what modalities interact stronger than others, and hidden cardiovascular dependency learned by the fusion module.

3.6.3 Aggregated Feature Importance Ranking

To support clinician readability, the aggregate importance for each clinical feature is computed by averaging SHAP scores across all classes:

$$R_j = \frac{1}{3}\sum_{c=1}^{3} I_j^{(c)} \quad (49)$$

The ranked list of features $R = \{R_1, R_2, \dots, R_M\}$ is ordered by ranking in descending order to discover the most important pathology indicators at patient-level that contribute to model inference. This allows physicians to view the main cardiovascular biomarkers identified for automated screening. The attention-driven intrinsic importance learned by the Attention-Guided Clinical Encoder is depicted in Figure 4.

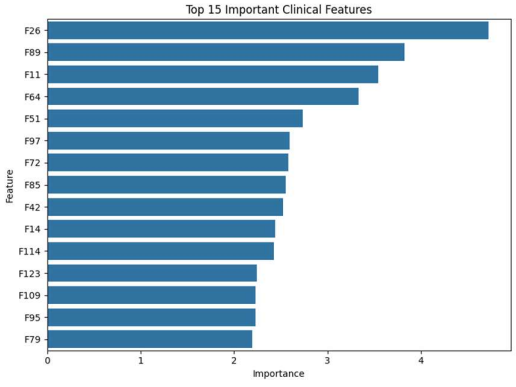


Figure 4. Attention-derived intrinsic feature importance obtained from the Attention-Guided Clinical Encoder

The proposed Explainable Decision Interpretation module offers Global transparency, Class-wise transparency, Fusion transparency, and Clinician trust enhancement. Therefore, the proposed hybrid framework XCardioNet is enhanced through the combination of SHAP attribution and cross-modal attention visualization making it more interpretable, reliable, and practical for real-world cardiovascular screening applications.

3.7 GUI-Based Real-Time Diagnostic Deployment

In addition, a smart Graphical User Interface (GUI) for cardiovascular diagnosis in real time is developed to further validate the proposed XCardioNet framework. The GUI is carefully structured in the 2 × 2 bordered layouts, which can intuitively visualize multimodal inputs, processing processes and final predictions. The buttons in the upper-left panel control the workflow: Load MMData, Preprocess, Extract Features, Classify and Exit; they are enabled one after another ensuring a correct diagnostic pipeline. Upper right: Data entry features for structured clinical record of patients; Lower left: corresponding ECG waveform plots for electrophysiological visualization.

When executing a prediction, the ECG and clinical data that are independently preprocessed and passed through their respective trained feature encoders, after which the embeddings extracted at this level will be tied together by multimodal classifier for contemporaneous cardiovascular decision. The lower-right panel indicates the final predicted class (namely Healthy, Heart_Disease, and Heart_Attack), which provides patient-level classification at a glance. Furthermore, it validates the proposed framework's usability through interactivity without delay, memory economy of processing one sample at a time and stable inference. The interface thus developed is a useful prototype for intelligent clinical decision-support in real time cardiovascular screening. Figures 5 to 12 illustrate the various stages of the XCardioNet framework in predicting the result for the selected patient using multimodal data.

Figure 5. GUI Layout of the proposed XCardioNet framework

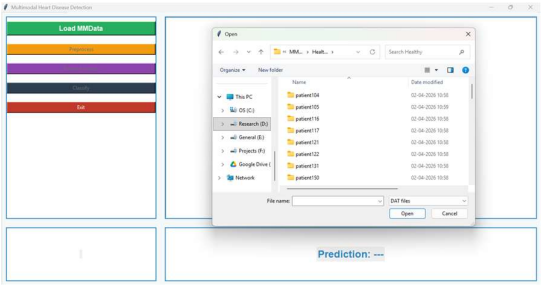


Figure 6. Interface of the ECG file selection dialog

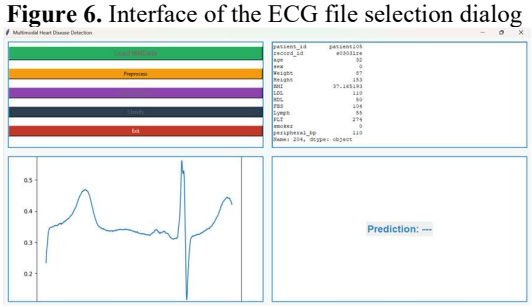


Figure 7. Visualization of the selected patient's ECG signal and clinical data in the GUI

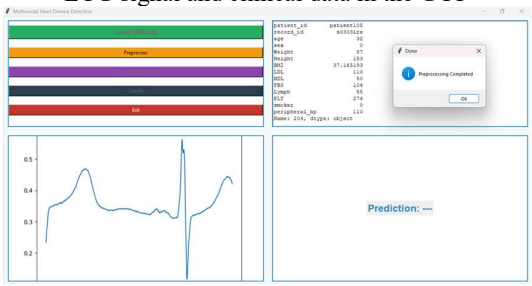


Figure 8. Display of a confirmation message after completion of preprocessing in the GUI.

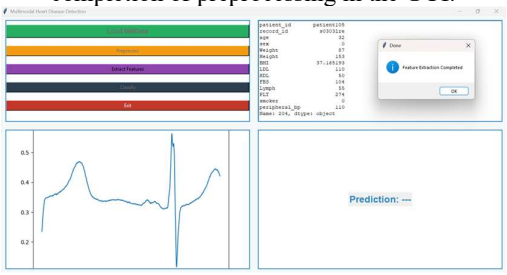


Figure 9. Display of a confirmation message after completion of feature extraction in the GUI

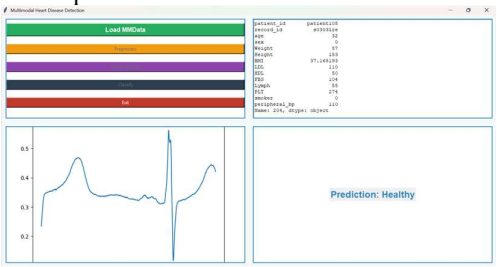


Figure 10. The GUI's result window shows the predicted class 'Healthy' for the selected patient

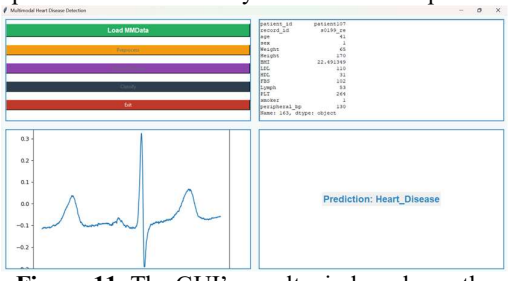


Figure 11. The GUI's result window shows the predicted class 'Heart_Disease' for the selected patient

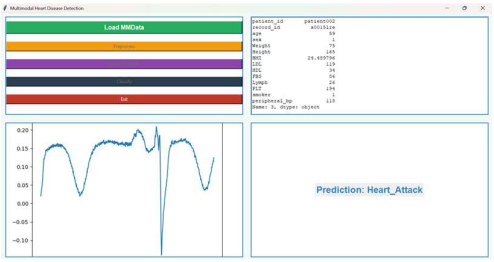


Figure 12. The GUI's result window shows the predicted class 'Heart_Attack' for the selected patient

The overall proposed model framework is described in the following algorithm.

Algorithm: Explainable Multimodal Cardiovascular Disorder Diagnosis via XCardioNet Cross-Modal Attention Fusion of ECG and Clinical Features
Input: Paired multimodal dataset of ECG waveform records and structured clinical attributes. **Output:** Trained tri-class classifier (Healthy, Heart_Disease, Heart_Attack), evaluation metrics, SHAP explanations, attention heatmaps, and GUI-based real-time testing workflow.
1. Initialize Components a. Define modality-specific encoders and fusion head: • ECG encoder: CNN + BiLSTM + Transformer to produce 128-D electrophysiological embedding. • Clinical encoder: Dense projection + attention weighting to produce 128-D pathological embedding. • Cross-Modal Fusion: 4-head Multihead Attention in a shared 128-D latent space. • Tri-Class Classification Head: 128→64 (ReLU) → Dropout(0.3) → 64→3 (Softmax). b. Set training parameters: batch size (64), learning rate (1e-3), optimizer (Adam), epochs (30). c. Create directories for feature storage, model checkpoints, and result output. 2. Preprocess Dataset a. ECG preprocessing: read WFDB signal, Butterworth filtering, wavelet denoising, normalization, R-peak detection, heartbeat segmentation. b. Clinical preprocessing: KNN imputation, Isolation Forest anomaly removal, RobustScaler normalization. 3. Feature Extraction a. Train ECG encoder and extract patient-level 128-D ECG embeddings. b. Train clinical encoder and extract attention-guided 128-D clinical embeddings. c. Save ECG_features.npy, Clinical_features.npy and corresponding labels.

4. Multimodal Alignment and Dataset Split
   a. Align clinical features with ECG heartbeat samples.
   b. Form paired multimodal tensors and split into train/test subsets.
5. Train XCardioNet Fusion Model
   a. Project ECG and clinical embeddings through linear latent mapping.
   b. Apply Multihead Cross-Modal Attention for fused 128-D multimodal representation.
   c. Predict tri-class probabilities through classifier head.
   d. Compute CrossEntropyLoss and optimize using Adam backpropagation.
   e. Save trained model as fusion_model.pth.
6. Evaluation
   a. Compute Accuracy, Precision, Recall, and F1-score.
   b. Generate confusion matrix and multiclass ROC curve.
   c. Save metrics and plots to ./Results.
7. Explainability
   a. Compute SHAP values for global and class-wise clinical feature importance.
   b. Generate top feature importance chart.
   c. Extract average cross-modal attention weights and render attention heatmap.
   d. Perform objective comparison and ablation study visualization.
8. GUI-based Real-Time Testing
   a. Provide a 2×2 interface with controls: Load MMData, Preprocess, Extract Features, Classify, Exit.
   b. Display patient clinical record, ECG waveform, and predicted cardiovascular label.
   c. Enforce sequential button-state workflow for reliable testing.
9. Output
   • Trained models: ecg_model.pth, clinical_model.pth, fusion_model.pth
   • Results: metrics, confusion matrix, ROC curve, SHAP summaries, attention heatmap
   • Deployment: GUI-based intelligent cardiovascular screening prototype

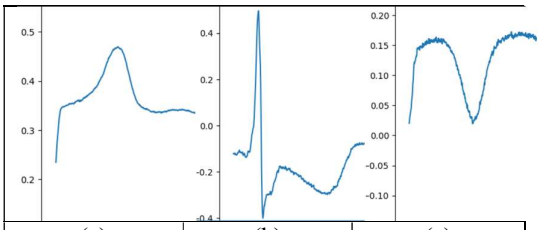
4. Results and Discussion

This research work presents an explainable multimodal cardiovascular disorder diagnosis framework in a hierarchical ECG representation learning, attention-guided clinical risk encoding and then multihead cross-modal attention fusion formulation collectively termed as the XCardioNet model. The developed framework is empirically validated on the publicly-available Multimodal Heart Health Dataset [24], which contains synchronized recordings of electrocardiogram (ECG) waveforms and structured patient clinical records for complete cardiovascular assessment. The multimodal dataset is chosen to show heterogeneous electrophysiological and pathological differences across three classes labelled as Healthy, Heart_Disease and Heart_Attack. In the experiments, all raw ECG signals are normalized to a cyclic and centered window of beats while their corresponding clinical attributes are converted into structured pathological feature vectors so that paired samples can be created consistently across modalities. To enable fair performance assessment, the multimodal dataset is split into training and testing subsets (80%:20%), stratified on all cardiovascular classes. During training, modality-specific encoders are first learned separately, following this the proposed XCardioNet fusion model is trained to perform synchronized tri-class cardiovascular classification under the same experimental conditions. With such a systematic experimental design, the assessment can be comprehensive regarding multimodal learning capability, cross-modal fusion effectiveness, explainability reliability and practical deployment feasibility. Table 2 shows an overview of the multimodal cardiovascular dataset used.

Table 2. Description of Multimodal Heart Health Dataset

Characteristics	Values
Dataset Source	PhysioNet
Modalities Used	ECG waveform + Clinical tabular attributes
No. of Diagnostic Classes	03
Class Labels	Healthy, Heart_Disease, Heart_Attack
ECG Record Format	WFDB (.dat, .hea)
ECG Signal Type	Multi-lead ECG
Sampling Frequency	360 Hz
Heartbeat Segments Used per Record	Maximum 50
Clinical Data Format	Structured CSV records
Clinical Attributes	Demographic + physiological + pathological indicators
Total Patient Records	277
Training–Testing Split	80:20 Stratified
Learning Objective	Tri-class cardiovascular disorder diagnosis

The sample ECG signal images of Healthy, Heart_Disease, and Heart_Attack are shown in Figure 13.



(a) (b) (c)

Figure 13. ECG signals of (a) Healthy (b) Heart_Disease (c) Heart_Attack

4.1 Objective-Wise Performance Assessment

In order to validate how each mentioned diagnostic objective contributes to the general effectiveness of the final XCardioNet framework, respectively independent experiments are performed based on unimodal clinical learning, unimodal ECG representation learning and finally on the multimodal cross-attention fusion strategy. The classification accuracies obtained are visualized in Figure 14. The clinical-only objective achieves an accuracy of 95.06% suggesting that patient pathological features convey informative and non-informative cardiovascular risk information but at the level of each alone only. The ECG-only objective also yields an increased accuracy of 96.79%, indicating that hierarchical electromagnetic physiogram learning captures stronger discriminative cardiac abnormalities compared to tabular pathology alone. In contrast, by jointly integrating both modalities with the proposed cross-modal attention fusion of evidence across data sources, the multimodal objective yields dramatically improved accuracy 99.13% thus revealing that synchronized physiological and clinical evidence offers considerably better cardiovascular diagnostic intelligence than isolated modality learning.

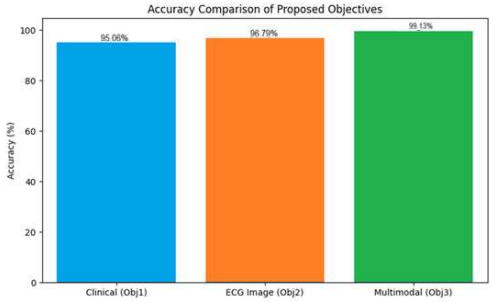


Figure 14. Objective-wise accuracy comparison of the proposed XCardioNet framework under Clinical-only, ECG-only, and Multimodal fusion learning settings

4.2 Overall Classification Performance of XCardioNet

The performance of the proposed model is then quantitatively assessed using four standard multiclass metrics (Accuracy, Precision, Recall and F1-score). The results indicate that XCardioNet can reach an overall accuracy of 99.13% with a precision of 99.15%, recall of 99.13%, and F1-score of 99.14%, demonstrating extremely well balanced classification consistency across all cardiovascular categories. The close agreement among all four metrics indicates that the model not only predicts correctly but also exhibits very low class bias and even lower false prediction propensity on them. This performance demonstrates the robustness of the proposed multimodal feature encoders and the discriminative power of the module with cross-attention fusion to handle heterogeneous cardiovascular evidence.

Table 3. Performance metrics of the Proposed XCardioNet Framework

Metrics	Results
Accuracy	99.13 %
Precision	99.15 %
Recall	99.13 %
F1-score	99.14 %

4.3 Confusion Matrix Analysis

The multiclass confusion matrix of the proposed XCardioNet model is depicted in Figure 15 to analyze prediction reliability for each class. The matrix shows that the prediction for Healthy is almost perfect with 800 correctly classified and very few wrongly classified. Similarly, the Heart_Disease class detects 409 true positives with minimal confusion towards the Healthy class. This is reflected in the fact that for the critical Heart_Attack category, 1545 samples are correctly recognized while only a very few confused with lower-risk categories. This assures that the multimodal framework would perform well even with one of the most severe acute cardiac abnormalities which is clinically vital for emergency decision support. These very sparse off-diagonal errors imply that the fused latent representation produced by XCardioNet has a high degree of inter-class separability.

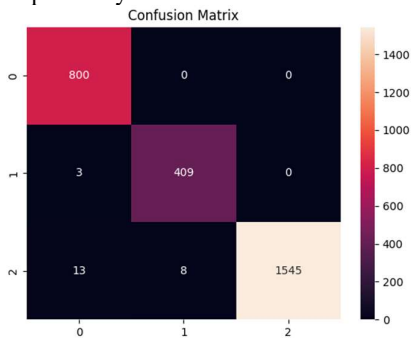


Figure 15. Confusion matrix of the proposed XCardioNet framework for tri-class cardiovascular disorder classification

4.4 ROC Curve and Discriminative Reliability

The class-wise discriminative confidence for the proposed framework is investigated using Receiver Operating Characteristic (ROC) analysis. All three cardiovascular classes show an area under curve (AUC) approaching 1·00 — indicating that the true positive versus false positive separation capability is

near perfect. As shown in the Figure 16, ROC trajectories are still very close to the upper-left corner, indicating that both multimodal latent embeddings from hierarchical ECG learner and clinical attention encoder result in a high posterior confidence estimation. The apparently ideal ROC behavior strongly supports using the proposed classifier in multiclass conditions for cardiovascular screening.

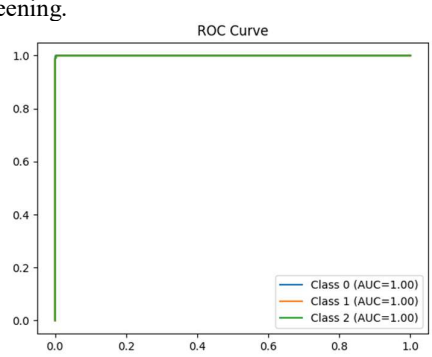


Figure 16. Multiclass ROC characteristics of the proposed XCardioNet model showing near-perfect AUC performance for Healthy, Heart_Disease, and Heart_Attack categories

4.5 Comparative Analysis with Existing State-of-the-Art Models

A comparative benchmark is also performed on three representative high-level cardiovascular diagnosis models, including HDPM [13], Explainable ECG-ML [21] and MAF-Net [22]. Results of the comparison are shown in Table 4. It shows that the clinical-only HDPM model results in markedly poorer accuracy compared to using electrophysiological information. The explainable model using ECG waveform morphology improves prediction robustness but still lacks patient pathological context. The performance of the multimodal MAF-Net architecture is still less hierarchical than the proposed design, although it derives strength from cross-attention fusion. XCardioNet, on the other hand, consistently performed best overall in all evaluation metrics using its built-in CNN–BiLSTM–Transformer ECG learner, risk-aware clinical attention encoder and adaptive multi-head multimodal fusion. It is thus established that the proposed framework provides a new state-of-the-art performance on the explainable cardiovascular disease diagnosis.

Table 4. Comparative performance analysis of the proposed XCardioNet framework against recent state-of-the-art cardiovascular diagnosis frameworks

Model	Input Modality	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
HDPM - Fitriyani et al. [13]	Clinical Only	94.82	94.57	94.63	94.49
Explainable ECG - ML - Majhi et al. [21]	ECG Only	96.41	96.18	96.22	96.09
MAF-Net - Qu et al. [22]	Multimodal Fusion	98.27	98.11	98.05	98.08
Proposed XCardioNet Framework	ECG + Clinical + Cross-Modal Attention	99.13	99.15	99.13	99.14

4.6 Ablation Study of Proposed Components

To analyze their individual contribution, we perform an ablation analysis by gradually enabling the key modules of our method and show the results in Figure 17. The clinical-only baseline of 95.06% and that of ECG-only learning improves the accuracy to 96.79%. The performance increases to 97.20% in the fusion of both modalities without attention guidance, demonstrating that heterogeneous evidence integration is beneficial. By including adaptive attention, it produces 97.80%, indicating the need for dependency-aware feature interaction. Lastly, the full XCardioNet architecture featuring hierarchical ECG learning, attention-guided clinical encoding, and multihead cross-modal fusion achieves the maximum accuracy of 99.13% in ablation experiments to demonstrate that all components cooperate cumulatively towards establishing the apex diagnostic superiority on ground truth labels.

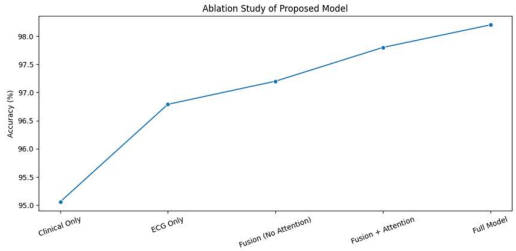


Figure 17. Ablation study demonstrating the progressive contribution of individual components in the proposed XCardioNet framework

5. Conclusion

This paper proposed an explainable hybrid deep learning framework: which is XCardioNet, which aims to accomplish the intelligent diagnosis of a tri-class cardiovascular disorder through joint utilization of electrocardiogram signals and structured clinical attributes. It unifies an ECG representation learner that combines hierarchical

CNN–BiLSTM–Transformer-based feature learning with attention-based pathological clinical implication encoding through a multihead cross-modal attention fusion mechanism to create synergy between electrophysiological waveform behavior and patient-centric cardiovascular risk evidence. Through a series of comprehensive experimental assessments, XCardioNet was demonstrated to achieve remarkably reliable class discrimination, outpacing its peers in terms of accuracy, precision, recall and F1-score with the support of complementary evidence from the confusion matrix, ROC characteristics as well as ablation analysis establishing strong potential for clinical applicability. The methodology also contributed to transforming the framework from a traditional black box predictor into clinically interpretable, and readily-implementable cardiovascular decision-support system with SHAP-based feature attribution and attention relevance interpretation integrated, support visualization by GUI-assisted real-time deployment. In summary, the results show that explainable multimodal deep attention fusion provides a generalizable and computationally efficient route to scalable next-generation precision cardiac screening, early abnormality identification, and risk-aware automated diagnosis for the general population. More future work may build upon the proposed framework by integrating more multimodal clinical evidence, including echocardiographic imaging and laboratory biomarkers, in addition to longitudinal patient monitoring signals blended with federated/cloud-assisted deployment strategies for large-scale hospital-level cardiovascular intelligence.

References

[1] Manimaran G, Chithra K, Karthikeyan R. Explainable deep learning based techniques for ECG-Based heart disease classification: A systematic literature review and future direction. *Computers in Biology and Medicine*. 2025; 199: 111324. https://www.sciencedirect.com/science/article/pii/S0010482525016786

[2] Zhang D, Yuan X, Zhang P. Interpretable Deep Learning for Automatic Diagnosis of 12-lead Electrocardiogram. *arXiv/Computational Cardiology Research*. 2020; 1-12. https://arxiv.org/abs/2010.10328

[3] Ahmad M, Almutairi F, Alghamdi A. Enhancing Heart Disease Diagnosis Using ECG Signal Reconstruction and Deep Transfer Learning Classification with Optional SVM Integration. *MDPI Diagnostics*. 2025; 15(12): 1501. https://www.mdpi.com/2075-4418/15/12/1501

[4] Poonkodi P, Chouhan M, Sobia MSC, Durairaj S. Detection of cardiovascular disease using deep learning-based attention-guided Bi-LSTM with electrocardiogram signals. *Intelligent Decision Technologies*. 2025; 19(3): 1-18. https://journals.sagepub.com/doi/10.1177/18724981251318209

[5] Cao TM, Tran NH, Nguyen PL, Pham H. Multimodal contrastive learning for diagnosing cardiovascular diseases from electrocardiography (ECG) signals and patient metadata. *arXiv/ Signal Processing*. 2023; 1-14. https://arxiv.org/abs/2304.11080

[6] Mohsen F, Soliman M, Hassan A. ECG features improve multimodal deep learning prediction of incident T2DM in a Middle Eastern cohort. *Springer Nature Scientific Reports*. 2025; 15: 27164. https://www.nature.com/articles/s41598-025-12633-z

[7] Gupta V, Hilgendorf L, Andersson E, Louca A, Shahmari A, Hjalmarsson A, Saini R, Pirazzi C, Alchay M, Rawshani A. Multimodal deep learning for acute myocardial infarction detection from 12-lead electrocardiogram: a multi-centre study with cross-hospital validation. *European Heart Journal – Digital Health*. 2025; 7(2): ztaf125. https://academic.oup.com/ehjdh/advance-article/doi/10.1093/ehjdh/ztaf125/8314874

[8] Wu S, Zhou J, Dong Y, Chen F. Enhancing Explainability of Deep Learning-Based ECG Diagnosis Using Large Language Models. *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence*. 2024; 61-65. https://dl.acm.org/doi/full/10.1145/3704137.3704146

[9] Lee C-C, Chuang C-C, Yeng C-H, So E-C, Chen Y-J. A cross-stage partial network and a cross-attention-based transformer for an electrocardiogram-based cardiovascular disease decision system. Bioengineering. 2024; 11(6): 549. https://www.mdpi.com/2306-5354/11/6/549

[10] Chang C-H, Lin C-S, Lee C-H, Lin C, Lee C-C, Liu W-T, Lee Y-T, Tsai D-J. Real-world application of deep learning for ECG-based prediction of coronary artery disease and revascularization needs. *European Heart Journal – Digital Health*. 2025; 6(6): 1124–1133. https://academic.oup.com/ehjdh/article/6/6/1124/8238774

[11] Dhandapani S, Somasundaram H, Angamuthu T. Hybrid deep learning framework for heart disease prediction using ECG signal images. *Scientific Reports*. 2025; 15: 33922. https://pmc.ncbi.nlm.nih.gov/articles/PMC12484791/

[12] Lilhore UK, Simaiya S, Khan M, Alroobaea R, Baqasah AM, Alsafyani M, Alhazmi A. A deep learning approach for heart disease detection using a modified multiclass attention mechanism with BiLSTM. *Scientific Reports*. 2025; 15: 25273. https://www.nature.com/articles/s41598-025-09594-8

[13] Fitriyani NL, Syafrudin M, Alfian G, Rhee J. HDPM: An effective heart disease prediction model

for a clinical decision support system. IEEE Access. 2020; 8: 133034–133050. https://ieeexplore.ieee.org/document/9144587/authors#authors

[14] Cai Y, Cai Y-Q, Tang L-Y, Wang Y-H, Gong M, Jing T-C, Li H-J, Li-Ling J, Hu W, Yin Z, Gong D-X, Zhang G-W. Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review. *BMC Medicine*. 2024; 22: 56. https://pmc.ncbi.nlm.nih.gov/articles/PMC10845808/

[15] Bartusik-Aebisher D, Rogóż K, Aebisher D. Artificial Intelligence and ECG: A New Frontier in Cardiac Diagnostics and Prevention. *Biomedicines*. 2025; 13(7):1685. https://pmc.ncbi.nlm.nih.gov/articles/PMC12292989/

[16] Sun S, Zhou C, Sun G, Lian Z. Multimodal fusion of ECG and chest X-ray with deep learning and radiomics for cardiac diagnosis: A machine learning-based nomogram approach. *Journal of Radiation Research and Applied Sciences*. 2025; 18(4): 102020. https://www.sciencedirect.com/science/article/pii/S1687850725007320

[17] Soto JT, Hughes JW, Sanchez PA, Perez M, Ouyang D, Ashley EA. Multimodal deep learning enhances diagnostic precision in left ventricular hypertrophy. *European Heart Journal – Digital Health*. 2022; 3(3): 380–389. https://academic.oup.com/ehjdh/article/3/3/380/6590817

[18] Yang X, Li Y, Wang J, Jia Y, Yi Z, Chen M. Utilizing multimodal artificial intelligence to advance cardiovascular diseases. *Precision Clinical Medicine*. 2025; 8(3): pbaf016. https://academic.oup.com/pcm/article/8/3/pbaf016/8205560

[19] Archana P, Shashikala SV. A Hybrid Deep Learning Heart Disease Prediction Framework Utilizing Multi-Modal Medical Imaging and Novel Feature Fusion Techniques. *Engineering, Technology & Applied Science Research*. 2025; 15(6): 28596-28602. https://etasr.com/index.php/ETASR/article/view/13204

[20] Ramos-Zaga FA. Artificial intelligence and multimodal diagnostic approaches in cardiovascular disease. *Archivos Peruanos de Cardiologia Y Cirugia Cardiovascular*. 2025; 6(4):230–238. https://pmc.ncbi.nlm.nih.gov/articles/PMC12825441/

[21] Majhi B, Kashyap A. Explainable AI-driven machine learning for heart disease detection using ECG signal. *Applied Soft Computing*. 2024; 167(A): 112225. https://www.sciencedirect.com/science/article/abs/pii/S1568494624009992

[22] Qu C, Zhang X, Lu Y, Wang Y, Su C. MAF-Net: Multimodal cross-attention-based fusion network for cardiovascular disease classification. PLoS One. 2026; 21(4):e0345238. https://pmc.ncbi.nlm.nih.gov/articles/PMC13056181/

[23] Dong R, Xie L. A multimodal cross-attention-based CNN-LSTM network for arrhythmia classification and clinical diagnostic support. *Biomedical Signal Processing and Control*. 2026; 112(D): 108697. https://www.sciencedirect.com/science/article/abs/pii/S174680942501208X

[24] Dataset Source [Online] Available: https://physionet.org/content/ptbdb/1.0.0/